

## IMPROVING THE CONVERGENCE OF REVERSIBLE SAMPLERS

LUC REY-BELLET

*Department of Mathematics and Statistics  
University of Massachusetts Amherst, Amherst, MA, 01003*

KONSTANTINOS SPILIOPOULOS

*Department of Mathematics and Statistics  
Boston University, Boston, MA, 02215*

**ABSTRACT.** In Monte-Carlo methods the Markov processes used to sample a given target distribution usually satisfy detailed balance, i.e. they are time-reversible. However, relatively recent results have demonstrated that appropriate reversible and irreversible perturbations can accelerate convergence to equilibrium. In this paper we present some general design principles which apply to general Markov processes. Working with the generator of Markov processes, we prove that for some of the most commonly used performance criteria, i.e., spectral gap, asymptotic variance and large deviation functionals, sampling is improved for appropriate reversible and irreversible perturbations of some initially given reversible sampler. Moreover we provide specific constructions for such reversible and irreversible perturbations for various commonly used Markov processes, such as Markov chains and diffusions. In the case of diffusions, we make the discussion more specific using the large deviations rate function as a measure of performance.

**Keywords:** Markov processes, Monte Carlo Sampling, Irreversibility, Detailed balance, Langevin Sampling, Large deviations, Asymptotic Variance

## 1. INTRODUCTION

In this paper we study the problem of sampling from a probability distribution  $\pi(dx)$  which, typically, is known only up to a normalizing constant. Sampling directly from  $\pi(dx)$  is often infeasible and thus one needs to rely on approximations. For example if  $f : E \mapsto \mathbb{R}$  is a given observable on the state space  $E$  and if one is interested in computing  $\bar{f} = \int_E f(x)\pi(dx)$  one constructs a positive recurrent Markov process  $X(t)$  which has  $\pi$  as its invariant distribution. Using the ergodic theorem

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t f(X(s))ds = \bar{f}, \text{ a.s. for } f \in L^1(\pi).$$

one can approximate  $\bar{f}$  by  $f_t = \frac{1}{t} \int_0^t f(X(s))ds$  for sufficiently large  $t$ . Clearly the degree to which such an approximation is efficient depends on the ergodic properties of the Markov process  $X(t)$  and on the criterion used for comparison.

Many different reversible and irreversible algorithms have been proposed in the literature dealing with both discrete and continuous (time or space) Markov chains as well as diffusion processes. For Markov chains we refer the reader to [2, 3, 4, 6, 7, 13, 14, 22, 26, 27, 28, 29, 30, 35, 37] and for diffusion processes we refer the reader to [9, 18, 19, 20, 21, 31, 32]. In most of these works, a reversible Markov chain or diffusion,  $X_0(t)$ , that has  $\pi$  as its invariant distribution is taken as a reference process and then different reversible or irreversible perturbations are explored which maintain the same invariant measure and lead to improved

---

*E-mail addresses:* luc@math.umass.edu, kspiliop@math.bu.edu.

*Date:* June 10, 2016.

K.S. was partially supported by the National Science Foundation (NSF) DMS 1312124 and during revisions of this article by NSF CAREER award DMS 1550918. LRB was partially supported by the NSF DMS 1109316.

sampling properties. The criteria that are mostly used for comparison purposes are the spectral gap of the generator of the process and the asymptotic variance of the estimator. Relatively recently, the large deviations rate function has been proposed in [10] and used in [10, 31, 32] as an alternative criterium for convergence and its connection to the asymptotic variance have been explored.

The contribution of this paper is threefold. Firstly, we unify and extend existing results in the literature demonstrating that there is a general underlying principle that applies to virtually all appropriate modifications of given reversible Markov processes, without having to restrict attention to continuous or discrete Markov jump processes or diffusion processes. Working directly with the infinitesimal generator of the Markov process, we prove that, under suitable conditions, reversible perturbations by negative definite generators as well as irreversible perturbations that maintain the invariant measure result in faster convergence to equilibrium. We prove that this is true based on all commonly used criteria of convergence; spectral gap, asymptotic variance and large deviations. We remark however that in this paper we restrict attention to additive perturbations of a generator by assymetric and anti-symmetric operators and we do not discuss techniques such as importance sampling, splitting, stratification and sequential sampling.

Secondly, we discuss specific constructions of such reversible and irreversible perturbations. We focus on continuous time Markov chains, Markov jump processes and diffusion processes. Some of these specific constructions are known in the literature, such as the Peskun and Tierney constructions, [30, 37], whereas others are novel, such as the reversible perturbation of Markov jump processes, Example 2, the reversible perturbation of diffusion processes, Example 3 and the irreversible perturbations of generic Markov chains, Example 4.

Thirdly, following [10, 31, 32] we argue that large deviations is a natural criterion for comparison for ergodic averages since it looks directly at the actual numerical approximation, which is the ergodic average. It has the advantage that in many cases it allows explicit computations which helps when comparing different algorithms. Also, it is directly connected to the asymptotic variance, as the second order Taylor expansion of the large deviation rate function around the limit  $\bar{f}$  is inversely proportional to the asymptotic variance. We focus mainly on diffusion processes where the known form of the rate function allows to do comparisons among specific algorithms.

The rest of the paper is organized as follows. In Section 2, we discuss the type of allowed perturbations and provide specific examples of such possible perturbations for cases of interest, such as Markov chains, Markov jump processes and diffusion processes. In Section 3 we prove that the previously mentioned perturbations lead to improvement of sampling based on the behavior of the spectral gap, the asymptotic variance and of the large deviations rate function. In Section 4 we study some consequences of our theory for irreversible perturbations of Markov Chains. Moreover, using large deviations, in Section 5, we study the effect of appropriate negative reversible and irreversible perturbations of reference reversible diffusion processes on the rate of convergence to equilibrium.

## 2. PERTURBATIONS OF REVERSIBLE MARKOV PROCESSES

Let us consider an ergodic time reversible continuous-time Markov process  $X_0(t)$  on the state space  $K$  with invariant measure  $\pi$ . Let  $L_{\mathbb{R}}^2(\pi)$  be the real Hilbert space with scalar product  $\langle f, g \rangle = \int f(x)g(x)\pi(dx)$ . We denote by  $T_t^0$  the corresponding strongly continuous Markov semigroup as an operator on  $L_{\mathbb{R}}^2(\pi)$  with infinitesimal generator  $\mathcal{L}_0$  with domain  $D(\mathcal{L}_0)$ . When discussing spectral properties we will need also to consider  $L_{\mathbb{C}}^2(\pi)$ , the complex Hilbert space with scalar product  $\langle f, g \rangle = \int f(x)\bar{g}(x)\pi(dx)$ . All operators involved here are real operators (they map real functions into real functions) and so they extend trivially to  $L_{\mathbb{C}}^2(\pi)$ . Abusing notation slightly, we will use the same notation for the operators acting on the real or complex Hilbert spaces.

Since  $X_0(t)$  is time-reversible,  $T_t^0$  and its generator  $\mathcal{L}_0$  are self-adjoint: that is we have

$$(1) \quad \langle f, \mathcal{L}_0 g \rangle = \langle \mathcal{L}_0 f, g \rangle$$

for all  $f, g \in D(\mathcal{L}_0)$ .

We shall also assume the semigroup  $T_t^0$  has a spectral gap in  $L_{\mathbb{R}}^2(\pi)$ , i.e., there exists  $\lambda_0 < 0$  such that

$$(2) \quad \sigma(\mathcal{L}_0) \setminus \{0\} \subset (-\infty, \lambda_0]$$

where  $\sigma(\mathcal{L}_0)$  denotes the spectrum. Note that  $\mathcal{L}_0$  is then negative definite, i.e. we have

$$(3) \quad \langle f, \mathcal{L}_0 f \rangle \leq 0$$

for all  $f$  in  $L^2_{\mathbb{C}}(\pi)$ .

We may think of  $X_0(t)$  as a reference process and we now introduce two types of “perturbations”  $X(t)$  of the processes  $X_0(t)$  where we require that  $X(t)$  has the same invariant measure  $\pi$ . In the first type of perturbation  $X(t)$  maintains the reversibility property, even though the dynamics have changed, whereas in the second type of perturbation  $X(t)$  is no longer reversible.

We describe then simple criteria which ensure that the process  $X(t)$  converges faster to equilibrium than  $X_0(t)$  in various senses. In Section 3 we prove that these perturbations lead to faster convergence to equilibrium. Essentially, reversible perturbations by negative definite operators and irreversible perturbations, that is perturbations by adding an antisymmetric part to the generator, lead to improvement in sampling.

In Subsection 2.1 we look at reversible perturbations and Examples 1-3 present some concrete constructions. Then, in Subsection 2.2 we look at irreversible perturbations and Examples 4-5 present some related exact constructions.

**2.1. Reversible perturbations.** We consider a Markov process with generator  $\mathcal{L} = \mathcal{L}_0 + \mathcal{S}$  and we assume that

- (i) We have  $D(\mathcal{L}_0) \subset D(\mathcal{S})$  and  $\mathcal{L}_0 + \mathcal{S}$  is the generator of a Markov process with invariant measure  $\pi$
- (ii) For all  $f, g \in D(\mathcal{S})$  we have

$$\langle f, \mathcal{S}g \rangle = \langle \mathcal{S}f, g \rangle$$

i.e.  $\mathcal{S}$  is self-adjoint. This implies that both  $\mathcal{L}$  and  $\mathcal{L}_0$  are self-adjoint.

- (iii)  $\mathcal{S}$  is *negative definite* i.e.,

$$\langle f, \mathcal{S}f \rangle \leq 0$$

for all  $f \in D(\mathcal{S})$ .

Let us see now some specific examples of Markov processes where the perturbation  $\mathcal{S}$  can be constructed.

**Example 1. (Markov chains on finite discrete spate space, Peskun condition).** Consider a continuous-time Markov chain on a discrete finite state space  $K = \{1, \dots, N\}$  with transition probability kernel  $k_0(i, j)$ . The perturbation  $\mathcal{S}$  is such that for all pairs  $i, j$  in  $\mathcal{S}$  with  $i \neq j$  we have for the new transition probability kernel

$$(4) \quad k(i, j) \geq k_0(i, j).$$

This means that the jump rate for  $X(t)$  is bigger than the jump rate for  $X_0(t)$  for any part of the state. Intuitively it means that the Markov chains spends less time in its current state and this should speed up the convergence. This condition was introduced in [30] for discrete-time Markov chain and in [22] for continuous time and shown to lead to decreased variance.

Let us discuss now how one can construct  $\mathcal{S}$  concretely. Since we require both  $X$  and  $X_0$  to have the same invariant measure clearly  $\mathcal{S}(i, j)$  and  $\mathcal{S}(j, i)$  are not independent. But, for two different pairs of states  $(i, j)$  and  $(i', j')$  we can choose  $\mathcal{S}(i, j)$  and  $\mathcal{S}(i', j')$  completely independently. Thus, we can write

$$\mathcal{S} = \sum_{1 \leq i < j \leq N} \mathcal{S}^{(i, j)}$$

where  $\mathcal{S}^{(i, j)}$  has the form

$$\begin{pmatrix} -\epsilon & \cdots & \epsilon \\ \vdots & \vdots & \vdots \\ \delta & \cdots & -\delta \end{pmatrix}$$

and  $\delta$  and  $\epsilon$  are non-negative and satisfy

$$(5) \quad \pi(i)\epsilon = \pi(j)\delta.$$

The entries in  $\mathcal{S}^{(i,j)}$  are all zeros apart from the  $i, j$  rows and columns where the indicated values are taken. Condition (5) ensures that  $\mathcal{S}$  is self-adjoint since for any  $f = (f(1), \dots, f(N))^T$  and  $g = (g(1), \dots, g(N))^T$  in  $L_{\mathbb{R}}^2(\pi)$  we have

$$\begin{aligned} \langle f, \mathcal{S}^{(i,j)} g \rangle &= \pi(i)f(i)(-\epsilon g(i) + \epsilon g(j)) - \pi(j)f(j)(-\delta g(i) + \delta g(j)) \\ &= -\epsilon \pi(i)(f(i) - f(j))(g(i) - g(j)) \end{aligned}$$

which is obviously symmetric in  $f$  and  $g$ . So,  $\mathcal{S}$  is self-adjoint on  $L_{\mathbb{R}}^2(\pi)$  and then also on  $L_{\mathbb{C}}^2(\pi)$ . This ensures that both processes  $X$  and  $X_0$  satisfies detailed balance with respect to  $\pi$ . In addition we have for any  $f = (f(1), \dots, f(N))^T \in L_{\mathbb{R}}^2(\pi)$

$$\langle f, \mathcal{S}^{(i,j)} f \rangle = -\epsilon \pi(i)(f(i) - f(j))^2 \leq 0$$

hence  $\mathcal{S}$  is negative definite since  $\epsilon$  and  $\delta$  are non-negative.

**Example 2. (General jump process).** Let us consider a continuous time Markov jump process that has bounded infinitesimal generator taking values on a state space  $K$ . The general form of its generator takes the form

$$\mathcal{L}_0 f(x) = \lambda(x) \int_K (f(y) - f(x)) \alpha(x, dy)$$

where  $\lambda$  is a nonnegative bounded intensity function on  $K$  and  $\alpha(x, \Gamma)$  is a transition kernel on  $K \times \mathcal{B}(K)$ .

The construction of such a jump process can be done as follows. Consider a Markov chain  $X_n$  on  $K$  with transition probability  $\alpha(x, \Gamma)$  and letting  $\tau_1, \tau_2, \dots$  be independent (between them and from  $X_n$  for every  $n \in \mathbb{N}$  as well) and exponentially distributed random variables with mean 1, define  $s_\kappa$  via the relation  $\lambda(X_{\kappa-1})s_\kappa = \tau_\kappa$ . Then, the Markov jump process with generator  $\mathcal{L}_0$  is given by

$$(6) \quad X_0(t) = X_n, \quad \text{for} \quad \sum_{\kappa=1}^n s_\kappa \leq t < \sum_{\kappa=1}^{n+1} s_\kappa.$$

Let us assume that there exist  $0 < \lambda_1 \leq \lambda_2 < \infty$  such that for all  $x$ ,  $\lambda_1 \leq \lambda(x) \leq \lambda_2$ . Then, under appropriate conditions on the transition kernel  $\alpha$ , see for example Section 2 of [11], we have that  $X_0(t)$  is an ergodic process. In particular,  $\alpha$  has then an invariant distribution denoted by  $\tilde{\pi}$  and the boundedness of  $\lambda(\cdot)$  allows us to define

$$\pi(E) = \frac{\int_E \frac{1}{\lambda(x)} \tilde{\pi}(dx)}{\int_K \frac{1}{\lambda(x)} \tilde{\pi}(dx)}$$

which can be shown to be the unique invariant distribution of  $X_0(t)$ . Now, we also make the assumption that the process  $X_0$  is reversible, which means that for all  $x, y \in K$

$$(7) \quad \lambda(x)\alpha(x, dy)\pi(dx) = \lambda(y)\alpha(y, dx)\pi(dy).$$

There are many different reversible perturbations that one can imagine. Perhaps the simplest one is to use the Peskun-Tierney [30, 37] construction on the Markov chain  $X_n$  that is used to define the jump Markov process  $X_0(t)$  via (6), as follows. Notice that we can write

$$\mathcal{L}_0 f(x) = \int_K f(y) A(x, dy)$$

where setting  $\|\lambda\| = \sup_{x \in K} \lambda(x) > 0$ , we have defined

$$A(x, dy) = \nu \|\lambda\| (\hat{\alpha}(x, dy) - \delta_x(dy)), \quad \text{and} \quad \hat{\alpha}(x, dy) = \frac{\lambda(x)}{\|\lambda\|} \alpha(x, dy) + \left(1 - \frac{\lambda(x)}{\|\lambda\|}\right) \delta_x(dy).$$

Let us now consider a transition probability operator  $\beta(x, dy)$  such that for almost every  $x \in K$ ,  $\beta(x, \Gamma \setminus \{x\}) \geq \alpha(x, \Gamma \setminus \{x\})$  for every  $\Gamma \in \mathcal{B}(K)$ . Assume that  $\beta(x, dy)$  is such that (7) holds and consider the jump Markov process with generator

$$\mathcal{L} f(x) = \int_K f(y) B(x, dy)$$

where  $B(x, dy)$  is as  $A(x, dy)$  with  $\beta(x, dy)$  in place of  $\alpha(x, dy)$ .

Now, we are in the set-up of [37]. It is easy to see that for almost every  $x \in K$  we have that  $B(x, \Gamma \setminus \{x\}) \geq A(x, \Gamma \setminus \{x\})$  for every  $\Gamma \in \mathcal{B}(K)$ . Lemma 3 in [37] guarantees that the operator  $\mathcal{L} - \mathcal{L}_0 = \mathcal{S}$  is negative

operator in  $\mathcal{L}^2(\pi)$ . In Section 3 we prove that if one uses the jump Markov process with generator  $\mathcal{L}f(x)$  instead of  $\mathcal{L}_0f(x)$ , then the sampling properties of the algorithm are better.

**Example 3. (Diffusions with multiplicative noise).** Let  $T > 0$  and consider the diffusion on  $\mathbb{R}^d$

$$(8) \quad dX(t) = [-\Sigma(X(t))\nabla U(X(t)) + T\nabla \cdot \Sigma(X(t))]dt + \sqrt{2T}\sigma(X(t))dB(t)$$

where  $B$  is a  $d$ -dimensional Brownian motion,  $\sigma : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ ,  $\Sigma(x) = \sigma(x)\sigma(x)^T$  and  $\nabla \cdot \Sigma$  denotes the vector field with components  $\sum_j \partial_{x_j} \Sigma_{i,j}(x)$ .

If  $\sigma$  is the identity matrix the equation reduces to the standard overdamped Langevin equation

$$(9) \quad dX_0(t) = -\nabla U(X_0(t))dt + \sqrt{2T}dB(t)$$

which we take as our reference process. The generator  $\mathcal{L}$  is given by

$$\mathcal{L} = T\nabla \cdot \Sigma \nabla - \Sigma \nabla U \cdot \nabla$$

In any case under suitable regularity and growth conditions on  $U$  and  $\sigma$  the process  $X(t)$  is ergodic and the measure

$$\pi(dx) = Z^{-1}e^{-U(x)/T}dx, \quad \text{with } Z = \int e^{-U(x)/T}dx$$

is invariant for (8) and  $X(t)$  is reversible. We have for  $f, g \in D(\mathcal{L})$

$$\langle f, \mathcal{L}g \rangle = -T \int \nabla f(x) \cdot \Sigma(x) \nabla g(x) \pi(dx).$$

Given that for  $f, g \in D(\mathcal{L}_0)$  the reference generator  $\mathcal{L}_0$  satisfies

$$\langle f, \mathcal{L}_0g \rangle = -T \int \nabla f(x) \cdot \nabla g(x) \pi(dx).$$

we get that the perturbation  $\mathcal{S}$  has the form

$$\langle f, \mathcal{S}g \rangle = - \int \nabla f(x) \cdot (\Sigma(x) - \mathbf{1}) \nabla g(x) \pi(dx).$$

A convenient choice is to take

$$\sigma(x) = \mathbf{1} + A(x)$$

where we choose  $A$  such that  $A + A^T$  is nonnegative definite. Then we have

$$\Sigma = \mathbf{1} + A + A^T + AA^T$$

and  $\mathcal{S}$  is negative definite. In the context of Hamiltonian Monte Carlo, the authors in [16] suggest using (8) with a special choice for the matrix  $\Sigma(x)$ . In Section 3 we prove that regular enough choices of  $\Sigma(x)$  such that  $\Sigma(x) - I$  is positive definite, lead to improved sampling. The degree of improvement depends of course on the choice of  $\Sigma(x)$ .

**2.2. Irreversible perturbations.** We consider a Markov process with generator  $\mathcal{L} = \mathcal{L}_0 + \mathcal{A}$  and we assume that

- (i) We have  $D(\mathcal{L}) \subset D(\mathcal{A})$  and  $\mathcal{L} + \mathcal{A}$  is the generator of a Markov process with invariant measure  $\pi$ .
- (ii) For all  $f, g \in D(\mathcal{A})$  we have

$$\langle f, \mathcal{A}g \rangle = -\langle \mathcal{A}f, g \rangle$$

i.e.  $\mathcal{A}$  is antiself-adjoint. Clearly this implies that

$$\langle f, \mathcal{A}f \rangle = 0$$

for all (real-valued)  $f \in D(\mathcal{A})$ .

**Example 4. (Markov chains on discrete state space).** Consider a continuous-time Markov chain on a discrete finite state space  $K = \{1, \dots, N\}$  with generator  $\mathcal{L}_0(i, j)$ . Comparisons of reversible and non-reversible Markov chains can be found in [3, 4, 6, 28, 29]. Here we present a simple irreversible perturbation of a reversible Markov chain that leads to acceleration of convergence.

To construct a non-reversible perturbation consider a matrix  $\Gamma(i, j)$  with

$$\Gamma(i, j) = -\Gamma(j, i), \quad \sum_j \Gamma(i, j) = 0$$

that is  $\Gamma$  is antisymmetric and the sum of its rows (and columns) is 0. Then set

$$\mathcal{A}(i, j) = \frac{1}{\pi(i)} \Gamma(i, j).$$

We have then

$$\sum_i \pi(i) \mathcal{A}(i, j) = \sum_i \Gamma(i, j) = 0$$

and this ensures that  $\pi$  is the invariant measure for the generator  $\mathcal{L} = \mathcal{L}_0 + \mathcal{A}$ . Of course one needs to choose the entries in  $\Gamma$  sufficiently small such that the entries in  $\mathcal{L}$  are nonnegative. Moreover, the adjoint of  $\mathcal{A}$  on  $L^2_{\mathbb{R}}(\pi)$  is the matrix with entries  $\mathcal{A}^*(j, i) = \pi(i) \mathcal{A}(i, j) \pi(j)^{-1}$  so that

$$\mathcal{A}^*(j, i) = \pi(i) \mathcal{A}(i, j) \pi(j)^{-1} = \Gamma(i, j) \pi(j)^{-1} = -\pi(j)^{-1} \Gamma(j, i) = -\mathcal{A}(j, i)$$

and thus  $\mathcal{A}$  is anti-selfadjoint.

To build concrete examples of such Markov chains we will express the perturbations in terms of *cycles*. To the reversible Markov chain with generator  $\mathcal{L}_0$  we associate, in the usual manner, the undirected graph  $G = (V, E)$  where the set of vertices  $V = K$  and where the edge  $(i, j)$  is in  $E$  if  $\mathcal{L}_0(i, j) > 0$ . Now we can construct irreversible perturbations in terms of *cycles* in the graph  $G$ . If we assume for example that the graph contains a cycle of length 3, say, through the states  $i, j, k$  in  $S$ , then we can take a perturbation  $\Gamma^{(i, j, k)}$  to be of the form

$$\Gamma^{(i, j, k)} \propto \begin{pmatrix} 0 & \dots & 1 & \dots & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & \dots & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \dots & -1 & \dots & 0 \end{pmatrix}$$

where the  $\dots$  and  $\vdots$  represent zero's and the elements shown are the  $i, j, k$  rows and columns. The proportionality constant must be chosen small enough so that the transition rates are non-negative. Then  $\mathcal{A} = \mathcal{A}^{(i, j, k)}$  has the form

$$\mathcal{A}^{(i, j, k)} \propto \begin{pmatrix} 0 & \dots & \frac{1}{\pi(i)} & \dots & -\frac{1}{\pi(i)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{1}{\pi(j)} & \dots & 0 & \dots & \frac{1}{\pi(j)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\pi(k)} & \dots & -\frac{1}{\pi(k)} & \dots & 0 \end{pmatrix}$$

To show how it can be achieved in a concrete Monte-Carlo situation consider the invariant measure  $\pi(i) = Z^{-1} e^{-H(i)}$ : any generator of the form

$$\mathcal{L}(i, j) = \frac{1}{\pi(i)} c(i, j) \text{ with } c(i, j) = c(j, i)$$

is reversible with invariant measure  $\pi(i)$ . Standard choices are the Glauber dynamics  $\mathcal{L}_G$  and the Metropolis dynamics  $\mathcal{L}_M$  with

$$\mathcal{L}_G(i, j) = \frac{e^{H(i)}}{e^{H(i)} + e^{H(j)}}, \quad \mathcal{L}_M(i, j) = e^{H(i)} \min \left\{ e^{-H(i)}, e^{-H(j)} \right\}$$

which both do not depend on the, possibly hard to compute, normalization constant  $Z$ .

One easily constructs adapted irreversible perturbations which in turn do not depend on the normalization constant by choosing for example for the Glauber dynamics

$$\mathcal{A}_G^{(i,j,k)} = \epsilon \begin{pmatrix} 0 & \cdots & \frac{e^{H(i)}}{e^{H(i)} + e^{H(j)} + e^{H(k)}} & \cdots & -\frac{e^{H(i)}}{e^{H(i)} + e^{H(j)} + e^{H(k)}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{e^{H(j)}}{e^{H(i)} + e^{H(j)} + e^{H(k)}} & \cdots & 0 & \cdots & \frac{e^{-H(j)}}{e^{H(i)} + e^{H(j)} + e^{H(k)}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{e^{H(k)}}{e^{H(i)} + e^{H(j)} + e^{H(k)}} & \cdots & -\frac{e^{H(k)}}{e^{H(i)} + e^{H(j)} + e^{H(k)}} & \cdots & 0 \end{pmatrix}$$

and with the abbreviation  $m(i, j, k) = \min \{e^{-H(i)}, e^{-H(j)}, e^{-H(k)}\}$  for the Metropolis dynamics

$$\mathcal{A}_M^{(i,j,k)} = \epsilon \begin{pmatrix} 0 & \cdots & e^{H(i)}m(i, j, k) & \cdots & -e^{H(i)}m(i, j, k) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -e^{H(j)}m(i, j, k) & \cdots & 0 & \cdots & e^{H(j)}m(i, j, k) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ e^{H(k)}m(i, j, k) & \cdots & -e^{H(k)}m(i, j, k) & \cdots & 0 \end{pmatrix}$$

Note that if we add perturbation for exactly one cycle we should take the coefficient  $\epsilon$  sufficiently small so that the rates are non-negative. In general if we add perturbations for many, or all, cycles the sum of the coefficients of all cycles containing any given state should not add up to more than 1 to ensure that the rates are non-negative. Also this can be generalized easily to cycles of arbitrary length and the details are left to the reader.

**Example 5. (Diffusions in  $\mathbb{R}^d$  or on compact manifolds)** Consider the SDE

$$dX(t) = -\nabla U(X(t)) + C(X(t)) + \sqrt{2T}dB(t)$$

in  $\mathbb{R}^d$  or in a compact manifold  $E$ . Assume that the growth properties of  $U(x)$  and  $C(x)$  are such that the SDE has a unique, non-explosive strong solution with a unique invariant measure. For  $C(x) = 0$  the process is reversible with invariant measure

$$\pi(dx) = Z^{-1} e^{-U(x)/T} dx, \quad \text{where } Z = \int e^{-U(x)/T} dx$$

and generator

$$\mathcal{L}_0 = T\Delta - \nabla U \cdot \nabla$$

and if we pick  $C$  such that  $\text{div}(C(x)e^{-U(x)/T}) = 0$ , then the invariant measure  $\pi$  is maintained. Notice that

$$\mathcal{L} = \mathcal{L}_0 + \mathcal{A}, \quad \text{where } \mathcal{A} = C \cdot \nabla$$

and then  $\mathcal{A}$  is antisymmetric in  $L^2(\pi)$ . The relation  $\text{div}(C(x)e^{-U(x)/T}) = 0$  is equivalent to  $\text{div}(C(x)) = T^{-1}C(x) \cdot \nabla U(x)$ , which is implied if we assume that  $C$  is divergence free and orthogonal to  $\nabla U$ , i.e.,  $\text{div}(C(x)) = 0$  and  $C(x) \cdot \nabla U(x) = 0$ . The results of [19], for spectral gap, and of [31, 32] for the asymptotic variance and large deviations rate function, as well as [20, 9] for the asymptotic variance, show that the convergence improves when the irreversible perturbation  $\mathcal{A}$  is introduced.



### 3. GENERAL THEORY ON IMPROVEMENT OF CONVERGENCE PROPERTIES

We prove simple lemmas showing that perturbations of reversible and irreversible types ameliorate the convergence properties of the algorithms for all commonly used criteria of convergence: spectral gap, asymptotic variance and large deviations rate function.

In specialized settings, versions of Lemmas 1 and 2 below have appeared in the literature before, see [4, 6, 7, 9, 13, 14, 18, 19, 22, 26, 27, 28, 29, 30, 37]. The novelty of Lemmas 1 and 2 is that working solely with the generator, we can prove in great generality (i.e., without restricting to specialized settings) that perturbations of general reversible Markov processes by negative reversible or irreversible generators decrease both spectral gap and asymptotic variance of the estimator.

Lemmas 3 and 4 state that the large deviations behavior is also improved. This is because the tail probability of the estimator being away from the true value decreases faster, yielding faster convergence to equilibrium. This was studied in detail in [31] for the specific case of irreversible perturbations of reversible diffusion processes, i.e., in the setup of Example 5. Here we prove that this is true for Markov processes in general, without having to restrict attention to diffusion processes. We work directly with the generator of a given Markov process.

In the sequel we consider a generator of the type

$$\mathcal{L} = \mathcal{L}_0 + \mathcal{S} + \mathcal{A}$$

where  $\mathcal{S}$  is a reversible perturbation and  $\mathcal{A}$  is an irreversible one.

It will be useful to introduce the space

$$H_{\mathbb{R}}^0 \equiv \{f \in L_{\mathbb{R}}^2(\pi); \int f d\pi = 0\}$$

which is the subspace of  $L_{\mathbb{R}}^2(\pi)$  which is orthogonal to the eigenspace corresponding to the eigenvalue 0 of  $\mathcal{L}$ . In particular  $H_{\mathbb{R}}^0$  is invariant under the semigroup  $T^t$ .

In Lemma 1 we prove that the spectral gap associated to a Markov process with generator  $\mathcal{L}$  is smaller than the spectral gap associated to a Markov process with generator  $\mathcal{L}_0$ . In Lemma 2 we prove that the asymptotic variance of the empirical average of a Markov process improves (i.e., decreases) under the reversible and irreversible perturbation. Lastly, in Lemmas 3 and 4 we prove that a similar behavior is true from the eyes of the large deviations rate function for the empirical average.

**3.1. Spectral gap.** Our first result is about the spectral gap which is defined in the general (non-reversible) case as

$$\lambda = \sup\{\operatorname{Re}(z); z \in \sigma(\mathcal{L}), z \neq 0\}.$$

By the Hille-Philips theorem, see Section 12.3 [17], the existence of a spectral gap (i.e.  $\lambda < 0$ ) implies a bound

$$\|T_t f - \int f d\pi\| \leq C e^{\lambda t} \|f - \int f d\pi\|$$

for all  $f \in L_{\mathbb{R}}^2(\pi)$ . Here  $\|\cdot\|$  is the  $L_{\mathbb{R}}^2(\pi)$  norm. Note that in the reversible case, i.e. when  $T_t = T_t^0$  is associated with  $\mathcal{L}_0$ , the spectral theorem implies that the constant  $C$  is equal to 1.

**Lemma 1. [Spectral gap].** *The spectral gap  $\lambda$  of the generator of semigroup with generator  $\mathcal{L} = \mathcal{L}_0 + \mathcal{S} + \mathcal{A}$  is smaller than the spectral gap  $\lambda_0$  of  $\mathcal{L}_0$ .*

*Proof.* We will use the fact that the reference operator  $\mathcal{L}_0$  is self-adjoint. Let  $f \in D(\mathcal{L}) \subset H_{\mathbb{R}}^0(\pi)$ . Using that for real-valued  $f$ ,  $(f, \mathcal{A}f) = 0$  we then obtain

$$\frac{d}{dt} \|T_t f\|^2 = 2\langle T_t f, (\mathcal{L}_0 + \mathcal{S} + \mathcal{A})T_t f \rangle = 2\langle T_t f, (\mathcal{L}_0 + \mathcal{S})T_t f \rangle \leq 2\langle T_t f, \mathcal{L}_0 T_t f \rangle \leq -2\lambda_0 \|T_t f\|^2.$$

The latter implies that  $\|T_t f\| \leq e^{-\lambda_0 t} \|f\|$  for any  $f \in H_{\mathbb{R}}^0$ . Notice that the pre-factor turns out to be  $C = 1$  here as well. But the norm of the real operator  $T_t f$  acting  $H_{\mathbb{R}}^0$  is the same as the norm on  $H_{\mathbb{C}}^0$ . Since  $\mathcal{L} + \lambda_0$  generates a contraction semigroup, we have that  $\operatorname{Re}(\langle (\mathcal{L} + \lambda_0)f, f \rangle) \leq 0$ , and so by Hille-Philips theorem we conclude that the spectrum of  $T_t$  lies in the half-plane  $\{\operatorname{Re}(z) \leq -\lambda_0\}$ .  $\square$



**3.2. Asymptotic variance.** We next turn to the asymptotic variance. Let  $f \in L^2_{\mathbb{R}}(\pi)$  be an observable and let  $\bar{f} = \int f d\pi$ . Note that  $f - \bar{f} \in H^0_{\mathbb{R}}$ . We assume that the operators  $\mathcal{L}_0$  and  $\mathcal{L}$  are invertible when restricted to  $H^0_{\mathbb{R}}$ . We denote by  $\mathcal{L}_0^{-1}$  and  $\mathcal{L}^{-1}$  their inverse which are bounded operators acting on  $H^0_{\mathbb{R}}$ .

For  $f \in L^2$  let  $S_t(f) = \int_0^t f(X(t)) dt$ , then  $\mathbf{E}_{\pi}(S_t(f)) = t\bar{f}$  and the asymptotic variance of  $S_t(f)/t$  satisfies

$$\sigma^2(f) \equiv \lim_{t \rightarrow \infty} \frac{1}{t} \text{Var}_{\pi}(S_t(f)/t) = 2 \int_0^{\infty} \langle T^t(f - \bar{f}), (f - \bar{f}) \rangle dt = \langle (f - \bar{f}), (-\mathcal{L})^{-1}(f - \bar{f}) \rangle$$

Using this we prove in Lemma 2 that the asymptotic variance never decreases by perturbations of the type  $\mathcal{S} + \mathcal{A}$ . Notice that in the case  $\mathcal{S} = 0$  a similar result has been recently obtained in [9] using different methods.

**Lemma 2. [Asymptotic variance].** *Let us assume that the operator  $(-\mathcal{L} - \mathcal{S})^{-1/2} \mathcal{A} (-\mathcal{L} - \mathcal{S})^{-1/2}$  is bounded. Then for any  $f \in L^2_{\mathbb{R}}(\pi)$  we have*

$$\sigma^2(f) \leq \sigma_0^2(f)$$

*Proof.* The reversible and irreversible perturbations use different arguments so we prove this in two steps. We first compare the variance for  $\mathcal{L}_0$  and  $\mathcal{L}_0 + \mathcal{S}$ . We can restrict ourselves on the subspaces  $H^0_{\mathbb{R}}$  where both operator are invertible. Since  $-\mathcal{L}_0$  is positive definite it possess a square root and we write

$$-\mathcal{L}_0 - \mathcal{S} = (-\mathcal{L}_0)^{1/2} \left( \mathbf{1} + (-\mathcal{L}_0)^{-1/2} (-\mathcal{S}) (-\mathcal{L}_0)^{-1/2} \right) (-\mathcal{L}_0)^{1/2}$$

and thus

$$(-\mathcal{L}_0 - \mathcal{S})^{-1} = (-\mathcal{L}_0)^{-1/2} \left( \mathbf{1} + (-\mathcal{L}_0)^{-1/2} (-\mathcal{S}) (-\mathcal{L}_0)^{-1/2} \right)^{-1} (-\mathcal{L}_0)^{-1/2}$$

By assumption  $-\mathcal{S}$  is non-negative definite therefore so is  $T = (-\mathcal{L}_0)^{-1/2} (-\mathcal{S}) (-\mathcal{L}_0)^{-1/2}$ . If we set

$$g = (-\mathcal{L}_0)^{1/2} (f - \bar{f})$$

the statement reduces to proving that for any  $g$  we have

$$\langle g, (\mathbf{1} + T)^{-1} g \rangle \leq \langle g, g \rangle$$

But this follows immediately from the spectral theorem for self-adjoint operator.

To handle the irreversible perturbation let us consider a generator of the form  $\mathcal{L}_0 + \mathcal{A}$  (if we have a symmetric perturbation  $\mathcal{S}$  replace  $\mathcal{L}_0$  by  $\mathcal{L}_0 + \mathcal{S}$ ). We notice first that any (bounded) operator  $B$  can be written as a sum of a self-adjoint part  $(B + B^*)/2$  and an anti self-adjoint part  $(B - B^*)/2$ . Since  $f$  is real-valued, only the self-adjoint part of the inverse of  $-\mathcal{L}_0 - \mathcal{A}$  matters in the asymptotic variance. To compute we first write

$$(-\mathcal{L}_0 - \mathcal{A})^{-1} = (-\mathcal{L}_0)^{-1/2} \left( \mathbf{1} + (-\mathcal{L}_0)^{-1/2} (-\mathcal{A}) (-\mathcal{L}_0)^{-1/2} \right)^{-1} (-\mathcal{L}_0)^{-1/2}$$

Set  $\mathcal{B} \equiv (-\mathcal{L}_0)^{-1/2} (-\mathcal{A}) (-\mathcal{L}_0)^{-1/2}$  which is anti-selfadjoint and thus  $(\mathbf{1} + \mathcal{B})(\mathbf{1} - \mathcal{B}) = \mathbf{1} - \mathcal{B}^2 = \mathbf{1} + \mathcal{B}^* \mathcal{B}$  we obtain

$$(\mathbf{1} + \mathcal{B})^{-1} = (\mathbf{1} + \mathcal{B}^* \mathcal{B})^{-1} - \mathcal{B}(\mathbf{1} + \mathcal{B}^* \mathcal{B})^{-1}.$$

Since  $\mathcal{B}^* = -\mathcal{B}$  and  $\mathcal{B}$  commutes with  $(\mathbf{1} + \mathcal{B}^* \mathcal{B})^{-1}$ , we then have that  $\langle g, \mathcal{B}(\mathbf{1} + \mathcal{B}^* \mathcal{B})^{-1} g \rangle = 0$ . The latter implies that

$$\sigma^2(f) = \langle (-\mathcal{L}_0)^{1/2} (f - \bar{f}), (\mathbf{1} + \mathcal{B}^* \mathcal{B})^{-1} (-\mathcal{L}_0)^{1/2} (f - \bar{f}) \rangle.$$

Since  $\mathcal{B}^* \mathcal{B}$  is nonnegative we conclude, as in the case of reversible perturbations, that  $\sigma_0(f) \leq \sigma(f)$ .  $\square$

**3.3. Large deviations.** Finally we turn to large deviations. Let us assume that all the processes involved satisfy a large deviation principle for the empirical measure

$$\mu_T = \frac{1}{T} \int_0^T \delta_{X(s)} ds$$

with a rate function  $I(\mu)$  which is given by Donsker-Varadhan formula

$$(10) \quad I(\mu) = - \inf_{u > 0, u \in D(\mathcal{L})} \int \frac{\mathcal{L}u}{u} d\mu.$$

Symbolically, we write

$$\mathbb{P} \{ \mu_t \approx \mu \} \asymp e^{-tI(\mu)}$$

where  $\asymp$  denotes logarithmic equivalence and the rate function  $I(\mu)$  quantifies the exponential rate at which the random measure  $\mu_t$  converges to  $\pi$ . Clearly, the larger  $I$  is, the faster the convergence occurs.

In particular, the formal definition is as follows. Let  $E$  be a Polish space, i.e., a complete and separable metric space. Denoting by  $\mathcal{P}(E)$  the space of all probability measures on  $E$ , we equip  $\mathcal{P}(E)$  with the topology of weak convergence, which makes  $\mathcal{P}(E)$  metrizable and a Polish space.

**Definition 1.** Consider a sequence of random probability measures  $\{\mu_t\}$ . The family  $\{\mu_t\}$  is said to satisfy a large deviations principle (LDP) with rate function (equivalently action functional)  $I : \mathcal{P}(E) \mapsto [0, \infty]$  if the following conditions hold:

- For all open sets  $O \subset \mathcal{P}(E)$ , we have

$$\liminf_{t \rightarrow \infty} \frac{1}{t} \log \mathbb{P} \{ \mu_t \in O \} \geq - \inf_{\mu \in O} I(\mu)$$

- For all closed sets  $F \subset \mathcal{P}(E)$ , we have

$$\limsup_{t \rightarrow \infty} \frac{1}{t} \log \mathbb{P} \{ \mu_t \in F \} \leq - \inf_{\mu \in F} I(\mu)$$

- The level sets  $\{\mu : I(\mu) \leq M\}$  are compact in  $\mathcal{P}(E)$  for all  $M < \infty$ .

We also assume that in the reversible case we have the following

$$(11) \quad I_o(\mu) = \left\langle \left( \frac{d\mu}{d\pi} \right)^{1/2}, (-\mathcal{L}_0) \left( \frac{d\mu}{d\pi} \right)^{1/2} \right\rangle$$

This has been proved in various cases, for discrete state space Markov chains and diffusions with smooth transition probability densities in [8]. The case of general jump processes is only partially understood, see [11] for the reversible case where a generalization of (11) is proved. Using this we obtain Lemma 3.

**Lemma 3. [Large deviations for empirical measures].** Let us consider measures  $\mu \in \mathcal{P}(E)$  such that  $(d\mu/d\pi)^{1/2} \in D(\mathcal{L}_0)$ . Let  $I(\mu)$  be the rate function associated with  $\mathcal{L}$  and  $I_o(\mu)$  the rate function associated with  $\mathcal{L}_0$ . We have

$$I(\mu) \geq I_o(\mu)$$

*Proof.* For reversible perturbations by negative definite operators  $\mathcal{S}$  this follows directly from the formula (11). For nonreversible perturbations, we can simply take  $u = u_0$  in (10), where

$$u_0 = \left( \frac{d\mu}{d\pi} \right)^{1/2}.$$

Then we have

$$I(\mu) \geq \int \frac{(-\mathcal{L}_0 - \mathcal{A})u_0}{u_0} d\mu = \int u_0(-\mathcal{L}_0 - \mathcal{A})u_0 d\pi = \int u_0(-\mathcal{L}_0)u_0 d\pi = I_o(\mu).$$

□

For  $f \in \mathcal{C}(E)$  the contraction principle implies that the ergodic average  $\frac{1}{t} \int_0^t f(X_s) ds$  satisfies a large deviation principle with rate function

$$\tilde{I}_f(\ell) = \inf_{\mu \in \mathcal{P}(E)} \{ I(\mu) : \langle f, \mu \rangle = \ell \}.$$

It is a nonnegative convex function with a minimum  $\tilde{I}_f(\bar{f}) = 0$  at  $\ell = \bar{f}$  and it is finite for the range of  $f$ , i.e. on the open interval  $(\min_x f(x), \max_x f(x))$ . One uses the informal notation  $\mathbb{P} \left\{ \frac{1}{t} \int_0^t f(X_s) ds \approx \ell \right\} \asymp e^{-t\tilde{I}_f(\ell)}$  to express that

$$\lim_{\epsilon \rightarrow 0} \lim_{t \rightarrow \infty} t \log \mathbb{P} \left\{ \frac{1}{t} \int_0^t f(X_s) ds \in (\ell - \epsilon, \ell + \epsilon) \right\} = \tilde{I}_f(\ell)$$

if  $\ell$  is in the range of  $f$ .

A Markov process whose rate function  $\tilde{I}_f(\ell)$  is higher means that its ergodic average converges faster to its equilibrium value. In fact, we have the following lemma.

**Lemma 4. [Large deviations for observables].** Consider  $f \in \mathcal{C}^{(\alpha)}(E)$  and  $\ell \in (\min_x f(x), \max_x f(x))$  with  $\ell \neq \int f d\pi$ . Then we have

$$\tilde{I}_f(\ell) \geq \tilde{I}_{f,0}(\ell),$$

where  $\tilde{I}_{f,0}(\ell) = \inf_{\mu \in \mathcal{P}(E)} \{I_o(\mu) : \langle f, \mu \rangle = \ell\}$ .

*Proof.* By definition  $\tilde{I}_f(\ell)$  is the infimum of  $I(\mu)$  over all  $\mu \in \mathcal{P}(E)$  such that  $\langle f, \mu \rangle = \ell$ . It easily follows by the affine form of the constraint  $\langle f, \mu \rangle = \ell$  in the definition of  $\tilde{I}_f(\ell)$ , that  $\tilde{I}_f(\ell)$  is a convex functional. Then, convexity and Lemma 3 trivially imply the statement of the lemma.  $\square$

We conclude this section by mentioning that Lemma 2 can be seen as a simple consequence of Lemma 4. Indeed, it is well known in the large deviations literature, see for example [5], that the asymptotic variance is inversely proportional to the second derivative of the large deviations rate function evaluated at  $\ell = \bar{f}$ , i.e.,

$$\sigma^2(f) = \frac{1}{2\tilde{I}_f''(\bar{f})}.$$

Then Lemma 4 and convexity of the rate function immediately imply the statement of Lemma 2. In addition to that, as we shall see in Section 5, a more careful analysis of the large deviations rate function reveals when there is a strict improvement in performance. It turns out that whether or not one has strict improvement in performance is related to the solution of a specific nonlinear Poisson equation. We note here that the Poisson equation that we derive is reminiscent of Poisson equations that have appeared in the literature in the analysis of MCMC algorithms, see for example Chapter 17 of [25]. In this paper, we see that the specific Poisson equation that we derive, characterizes when irreversible perturbations lead to strict improvement in performance.

#### 4. LARGE DEVIATIONS ANALYSIS OF IRREVERSIBLE PERTURBATION FOR MARKOV CHAINS.

We consider a finite state aperiodic irreducible Markov chain with transition probability kernel  $k_0(i, j)$  and invariant measure  $\pi$ . The generator of such a jump Markov process takes the form

$$\mathcal{L}_0 g(i) = \sum_j [k_0(i, j)(g(j) - g(i))]$$

Hence, the Donsker-Varadhan rate function takes the form

$$I(\mu) = - \inf_{g>0} \left[ \sum_i \frac{\mu(i)}{g(i)} \sum_j [k_0(i, j)(g(j) - g(i))] \right]$$

and as it is proven in [23] this can be simplified to

$$I(\mu) = \sum_{i,j} \mu(i) k_0(i, j) \left( 1 - e^{\frac{V_0(j) - V_0(i)}{2}} \right) = \sum_{i,j} \mu(i) k_0(i, j) - \sum_{i,j} \mu(i) k_0(i, j) e^{\frac{V_0(j) - V_0(i)}{2}}$$

where  $V_0$  is the unique solution (up to a constant) of the algebraic equation

$$(12) \quad \sum_j \left[ k_0(i, j) e^{\frac{V_0(j) - V_0(i)}{2}} \mu(i) - k_0(j, i) e^{\frac{V_0(i) - V_0(j)}{2}} \mu(j) \right] = 0, \text{ for all } i \in K.$$

The last relation shows that  $\kappa_{V_0}(i, j) = k_0(i, j) e^{\frac{V_0(j) - V_0(i)}{2}}$  is the transition probability density function for the Markov chain with invariant measure  $\mu$ . Obviously if  $\mu = \pi$ , the only possible solution to (12) is  $V_0(i) = \text{constant}$  for every  $i \in K$ . As expected, this of course means that  $I(\pi) = 0$ . Notice that under irreducibility, we can write

$$I(\mu) = \sum_{i,j} \mu(i) k_0(i, j) - \sum_{i,j} \mu(i) \kappa_{V_0}(i, j)$$

which means that the rate function can be viewed as the difference between the expected escape rates  $\sum_{i,j} \mu(i) k_0(i, j)$  and  $\sum_{i,j} \mu(i) \kappa_{V_0}(i, j)$ . The latter naturally estimates the difference in the number of transitions per unit time in the process.

As it has been observed in Example 4, if we consider a matrix  $\Gamma$  that is anti-symmetric and its rows sum to zero, i.e.,

$$\Gamma = -\Gamma^T \quad \text{and for every } i \in K \quad \sum_j \Gamma(i, j) = 0$$

then, the Markov chain with transition probability matrix  $k_\Gamma(i, j) = k_0(i, j) + \frac{1}{\pi(i)}\Gamma(i, j)$  will have the same invariant distribution  $\pi$ . Let us denote by  $V_\Gamma$  the solution to (12) with  $k_\Gamma$  in place of  $k_0$ .

Our goal is to compare the rate functions of the two Markov chains, the one with transition probability function  $k_0(i, j)$  and the one with transition probability function  $k_\Gamma(i, j)$ . Let us denote the associated large deviations rate functions by  $I_0(\mu)$  and  $I_\Gamma(\mu)$  respectively. Let us define, for a given transition rate function  $k(i, j)$  and a function  $V$  defined on the state space of the Markov chain, the functional

$$\mathcal{Y}_k(V) = \sum_{i,j} \mu(i)k(i, j)e^{\frac{V(j)-V(i)}{2}}$$

It is easy to see that the functional  $\mathcal{Y}_k(V)$  is non-negative and, under the irreducibility assumption, strictly convex, [23], with respect to functions  $V$  defined on the state space of the Markov chain. The unique minimum for  $\mathcal{Y}_{k_0}$  is attained at  $V = V_0$  whereas the unique minimum for  $\mathcal{Y}_{k_\Gamma}$  is attained at  $V = V_\Gamma$ .

Let us prove now, using Lemma 3 that  $\mathcal{Y}_{k_0}(V_0) \geq \mathcal{Y}_{k_\Gamma}(V_\Gamma)$ . In particular, this means that the minimum value of the functional  $\mathcal{Y}_{k_\Gamma}(\cdot)$  is below the minimum value of the functional  $\mathcal{Y}_{k_0}(\cdot)$ . This means that under irreversibility, there are more transitions per unit time in the process, which then naturally leads to faster convergence to equilibrium.

**Proposition 1.** *With the notation above we have that*

$$I_\Gamma(\mu) - I_0(\mu) = \mathcal{Y}_{k_0}(V_0) - \mathcal{Y}_{k_\Gamma}(V_\Gamma) \geq 0$$

*Proof of Proposition 1.* We have the following computations

$$\begin{aligned} I_\Gamma(\mu) - I_0(\mu) &= \sum_{i,j} \mu(i)k_\Gamma(i, j) \left(1 - e^{\frac{V_\Gamma(j)-V_\Gamma(i)}{2}}\right) - \sum_{i,j} \mu(i)k_0(i, j) \left(1 - e^{\frac{V_0(j)-V_0(i)}{2}}\right) \\ &= \sum_{i,j} \mu(i) \left(k_0(i, j) + \frac{1}{\pi(i)}\Gamma(i, j)\right) \left(1 - e^{\frac{V_\Gamma(j)-V_\Gamma(i)}{2}}\right) - \sum_{i,j} \mu(i)k_0(i, j) \left(1 - e^{\frac{V_0(j)-V_0(i)}{2}}\right) \\ &= \sum_{i,j} \mu(i) \left(k_0(i, j) + \frac{1}{\pi(i)}\Gamma(i, j) - k_0(i, j)\right) \\ &\quad + \left[ \sum_{i,j} \mu(i)k_0(i, j)e^{\frac{V_0(j)-V_0(i)}{2}} - \sum_{i,j} \mu(i) \left(k_0(i, j) + \frac{1}{\pi(i)}\Gamma(i, j)\right) e^{\frac{V_\Gamma(j)-V_\Gamma(i)}{2}} \right] \\ &= \sum_i \frac{\mu(i)}{\pi(i)} \sum_j \Gamma(i, j) + \\ &\quad + \left[ \sum_{i,j} \mu(i)k_0(i, j)e^{\frac{V_0(j)-V_0(i)}{2}} - \sum_{i,j} \mu(i) \left(k_0(i, j) + \frac{1}{\pi(i)}\Gamma(i, j)\right) e^{\frac{V_\Gamma(j)-V_\Gamma(i)}{2}} \right] \\ &= \left[ \sum_{i,j} \mu(i)k_0(i, j)e^{\frac{V_0(j)-V_0(i)}{2}} - \sum_{i,j} \mu(i) \left(k_0(i, j) + \frac{1}{\pi(i)}\Gamma(i, j)\right) e^{\frac{V_\Gamma(j)-V_\Gamma(i)}{2}} \right] \\ (13) \quad &= \mathcal{Y}_{k_0}(V_0) - \mathcal{Y}_{k_\Gamma}(V_\Gamma) \end{aligned}$$

In the last computation we used the fact that  $\sum_j \Gamma(i, j) = 0$ . Since, by Lemma 3, we have that  $I_\Gamma(\mu) \geq I_0(\mu)$ , we conclude the proof of the proposition.  $\square$

## 5. LARGE DEVIATIONS ANALYSIS OF REVERSIBLE AND IRREVERSIBLE PERTURBATION FOR DIFFUSIONS.

It turns out that, in the case of diffusion processes, the large deviations criterion can give more concrete information on how much improvement one gets by reversible and irreversible perturbations. Let us consider

the overdamped Langevin equation

$$(14) \quad dX_0(t) = -\nabla U(X_0(t))dt + \sqrt{2T}dB(t)$$

In Examples 3 and 5 we proposed specific reversible and irreversible perturbations of the infinitesimal generator of (9) that, based on Lemmas 1, 2, 3 and 4, lead to faster convergence to equilibrium, irrespectively of which performance criteria is being used. Our goal in this section is to characterize the improvement in sampling in more precise terms. We use the large deviations formalism for empirical measures.

As it turns out, we can write down how much the rate function increases when a reversible or an irreversible perturbation is performed. Based on the corresponding formula we can then characterize exactly when there is a strict increase in performance. The special case of irreversible perturbations of diffusions from Example 5 has been extensively studied in [19] based on spectral gap criteria and recently on [31, 32] based on the asymptotic variance and large deviations rate function criteria. We refer the interested reader to [19, 31, 32] for further details and for numerical results. In this section we compare how reversible and irreversible perturbations for general Markov processes compare via the lens of large deviations theory. The results of [19, 31, 32] are then essentially recovered as a special case of the general theory of this paper.

Let us start our analysis with a very general result on the large deviations principle for the invariant measure of diffusion processes. In order to avoid technical issues we shall restrict our discussion to diffusion taking values on a  $d$ -dimensional compact Riemannian manifold  $E$  of class  $C^3$  without boundary. In particular, we have the following general theorem.

**Theorem 1.** *Consider the SDE on  $E$  with infinitesimal generator*

$$\mathcal{L} = \frac{1}{2}\nabla \cdot a(x)\nabla + b(x)\nabla$$

*with  $b_i, a_{i,j} \in C^1(E)$ ,  $a(x)$  being strictly positive. Let  $\mu \in \mathcal{P}(E)$ , where  $\mu(dx) = p(x)dx$  is a measure with positive density  $p \in C^{(2+\alpha)}(E)$  for some  $\alpha > 0$ . The Donsker-Vardhan rate function  $I(\mu)$  takes the form*

$$(15) \quad I(\mu) = \frac{1}{8} \int_E \frac{\nabla p(x)a(x)\nabla p(x)}{p^2(x)} d\mu(x) - \frac{1}{2} \int_E \frac{b(x)\nabla p(x)}{p(x)} d\mu(x) + \frac{1}{2} \int_E \nabla \phi(x)a(x)\nabla \phi(x) d\mu(x)$$

*where  $\phi$  is the unique (up to constant) solution of the equation*

$$(16) \quad \text{div}[p(x)(b(x) + a(x)\nabla \phi(x))] = 0.$$

*Proof.* The proof of this theorem follows the same steps as that of Lemma 3.2 in [31] using the general results of Gärtner in [15]. Thus, the details are omitted.  $\square$

In the case of equation (9), i.e., when  $b(x) = -\nabla U(x)$  is a gradient and  $a(x) = 2TI$ , then  $\phi(x) = \frac{1}{2T}U(x) + \text{constant}$  and we get

$$(17) \quad I_o(\mu) = \frac{T}{4} \int_E \left| \frac{\nabla p(x)}{p(x)} + \frac{1}{T}\nabla U(x) \right|^2 d\mu(x)$$

which is the usual explicit formula for the rate function in the reversible case.

In this section we want to compare the rate function for the baseline case (9) with that of the reversible perturbation of Example 3 and that of the irreversible perturbation of Example 5.

For notational convenience, let us denote by

- (i)  $I_\Sigma(\mu)$  the rate function for the diffusion of Example 3, i.e., when  $a(x) = 2T\Sigma(x)$  and  $b(x) = -\Sigma(x)\nabla U(x)$ ,
- (ii)  $I_C(\mu)$  the rate function for the diffusion of Example 5, i.e., when  $a(x) = 2TI$  and  $b(x) = -\nabla U(x) + C(x)$ , and
- (iii)  $I_{\Sigma,C}(\mu)$  the rate function for the diffusion when  $a(x) = 2T\Sigma(x)$  and  $b(x) = -\Sigma(x)\nabla U(x) + C(x)$ .

Clearly, using this notation, the rate function for the reference case, (17), is  $I_o(\cdot) = I_{Id,o}(\cdot)$ .

Propositions 2, 3 and 4 summarize the increase of the Donsker-Varadhan rate functions for empirical measures based on reversible and irreversible perturbations. For presentation purposes, the proofs of these results is given at the end of the section. Moreover, based on these results we can then prove that the rate function for the empirical average of a given observable also increases under the suggested reversible and irreversible perturbations. This is Theorem 2. Conditions, guaranteeing strict improvement in performance

are also provided. It turns out that whether or not one has strict improvement in performance is related to the solution of a specific nonlinear Poisson equation.

**Proposition 2.** *Assume that the matrix  $\Sigma(x) \neq I$  is such that  $\Sigma(x) - I$  is nonnegative definite. For any  $\mu \in \mathcal{P}(E)$  we have  $I_\Sigma(\mu) \geq I_o(\mu)$ . If  $\mu(dx) = p(x)dx$  is a measure with positive density  $p \in \mathcal{C}^{(2+\alpha)}(E)$  for some  $\alpha > 0$  and  $\mu \neq \pi$  then we have*

$$I_\Sigma(\mu) - I_o(\mu) = \frac{T}{4} \int_E \left( \frac{\nabla p(x)}{p(x)} + \frac{1}{T} \nabla U(x) \right)^T (\Sigma(x) - I) \left( \frac{\nabla p(x)}{p(x)} + \frac{1}{T} \nabla U(x) \right) d\mu(x) \geq 0$$

Moreover we have that if  $p(x) > 0$  everywhere and  $\Sigma(x) - I$  is strictly positive everywhere, then  $I_\Sigma(\mu) > I_o(\mu)$ .

**Proposition 3.** *Assume that the vector field  $C(x) \neq 0$  is such that  $\text{div}(C(x)e^{-U(x)/T}) = 0$  and the matrix  $\Sigma(x)$  is strictly positive definite. For any  $\mu \in \mathcal{P}(E)$  we have  $I_{\Sigma,C}(\mu) \geq I_\Sigma(\mu)$ . If  $\mu(dx) = p(x)dx$  is a measure with positive density  $p \in \mathcal{C}^{(2+\alpha)}(E)$  for some  $\alpha > 0$  and  $\mu \neq \pi$  then we have*

$$I_{\Sigma,C}(\mu) - I_\Sigma(\mu) = 4T \int_E \left( \frac{1}{2} \nabla \phi(x) - \frac{1}{4T} \nabla U(x) \right)^T \Sigma(x) \left( \frac{1}{2} \nabla \phi(x) - \frac{1}{4T} \nabla U(x) \right) d\mu(x) \geq 0$$

where  $\phi$  is the unique solution (up to a constant) of the equation

$$\text{div}[p(x)(-\Sigma(x)\nabla U(x) + C(x) + 2T\Sigma(x)\nabla\phi(x))] = 0.$$

Moreover, if the positive density  $p(x)$  satisfies  $\text{div}(p(x)C(x)) \neq 0$ , then we have  $I_{\Sigma,C}(\mu) > I_\Sigma(\mu)$ . If  $p(x)$  is such that  $\text{div}(p(x)C(x)) = 0$ , then it has the form  $p(x) = e^{2G(x)}$  where  $G$  is such that  $G+U$  is an invariant quantity for the vector field  $C$  (i.e.,  $C\nabla(G+U) = 0$ ).

Clearly, if we set  $\Sigma(x) = I$ , then Proposition 3 shows that for the irreversible perturbation of Example 5 one has

$$I_C(\mu) - I_o(\mu) = 4T \int_E \left( \frac{1}{2} \nabla \phi(x) - \frac{1}{4T} \nabla U(x) \right)^T \left( \frac{1}{2} \nabla \phi(x) - \frac{1}{4T} \nabla U(x) \right) d\mu(x) \geq 0.$$

This is nothing else but Theorem 2.2 in [31]. As a matter of fact [31, 32] study in detail this special case via the lens of large deviations theory. We refer the interested reader to these articles for further details on this special case and related numerical simulation results. Next, in Proposition 4 we investigate the situation where one performs both reversible and irreversible perturbations.

**Proposition 4.** *Assume that the vector field  $C(x) \neq 0$  is such that  $\text{div}(C(x)e^{-U(x)/T}) = 0$  and the matrix  $\Sigma(x) - I$  is nonnegative definite. For any  $\mu \in \mathcal{P}(E)$  we have  $I_{\Sigma,C}(\mu) \geq I_o(\mu)$ . If  $\mu(dx) = p(x)dx$  is a measure with positive density  $p \in \mathcal{C}^{(2+\alpha)}(E)$  for some  $\alpha > 0$  and  $\mu \neq \pi$  then we have*

$$\begin{aligned} I_{\Sigma,C}(\mu) - I_o(\mu) &= \frac{T}{4} \int_E \left( \frac{\nabla p(x)}{p(x)} + \frac{1}{T} \nabla U(x) \right)^T (\Sigma(x) - I) \left( \frac{\nabla p(x)}{p(x)} + \frac{1}{T} \nabla U(x) \right) d\mu(x) \\ &\quad + 4T \int_E \left( \frac{1}{2} \nabla \phi(x) - \frac{1}{4T} \nabla U(x) \right)^T \Sigma(x) \left( \frac{1}{2} \nabla \phi(x) - \frac{1}{4T} \nabla U(x) \right) d\mu(x) \\ &\geq 0. \end{aligned}$$

where  $\phi$  is the unique solution (up to a constant) of the equation

$$\text{div}[p(x)(-\Sigma(x)\nabla U(x) + C(x) + 2T\Sigma(x)\nabla\phi(x))] = 0.$$

Moreover, if the positive density  $p(x)$  satisfies  $\text{div}(p(x)C(x)) \neq 0$  and  $\Sigma(x), \Sigma(x) - I$  are strictly positive definite, then we have  $I_{\Sigma,C}(\mu) > I_o(\mu)$ . If  $p(x)$  is such that  $\text{div}(p(x)C(x)) = 0$ , then it has the form  $p(x) = e^{2G(x)}$  where  $G$  is such that  $G+U$  is an invariant quantity for the vector field  $C$  (i.e.,  $C\nabla(G+U) = 0$ ).

Notice that the correction term in Proposition 4 is the sum of the correction terms from Propositions 2 and 3. This comes to no surprise, as the set-up of Proposition 4 is that of both reversible and irreversible perturbation.

Based on these results we then study the impact of these perturbations on the large deviations for the estimator  $f_t = \frac{1}{t} \int_0^t f(X_s) ds$  itself. For  $f \in \mathcal{C}(E)$  contraction principle implies that the ergodic average  $\frac{1}{t} \int_0^t f(X_s) ds$  satisfies a large deviation principle with rate function

$$\tilde{I}_f(\ell) = \inf_{\mu \in \mathcal{P}(E)} \{I(\mu) : \langle f, \mu \rangle = \ell\}.$$

As we remarked in Section 3, a Markov process whose rate function  $\tilde{I}_f(\ell)$  is higher means that its ergodic average converges faster to its equilibrium value, in the sense that the rate of the exponential convergence is faster.

By general principles, see for example [15], the rate function  $\tilde{I}_f(\ell)$  is given by the Legendre transform  $\tilde{I}_f(\ell) = \sup_{\beta \in \mathbb{R}} (\ell\beta - \lambda(\beta f))$  where

$$(18) \quad \lambda(\beta f) = \lim_{t \rightarrow \infty} \frac{1}{t} \log \mathbb{E}_x \left[ e^{\beta \int_0^t f(X_s) ds} \right]$$

Using a Perron-Frobenius argument one can show that  $\lambda(\beta f)$  is maximal eigenvalue of the operator  $\mathcal{L} + \beta f$  and that  $\lambda(\beta)$  is a smooth (real-analytic) function of  $\beta$  and hence

$$\tilde{I}_f(\ell) = \ell\hat{\beta} - \lambda(\hat{\beta}f)$$

where  $\hat{\beta} = \hat{\beta}(\ell)$  is the unique solution of  $\frac{d}{d\beta} \lambda(\beta f) = \ell$ .

We denote by  $\tilde{I}_{f,\Sigma,C}(\ell)$ ,  $\tilde{I}_{f,\Sigma}(\ell)$ ,  $\tilde{I}_{f,C}(\ell)$  and  $\tilde{I}_{f,o}(\ell)$  the rate functions corresponding to  $I_{\Sigma,C}(\mu)$ ,  $I_{\Sigma}(\mu)$ ,  $I_C(\mu)$  and  $I_o(\mu)$  respectively.

**Theorem 2.** Consider  $f \in \mathcal{C}^{(\alpha)}(E)$  and  $\ell \in (\min_x f(x), \max_x f(x))$  with  $\ell \neq \int f d\pi$ . Fix a vector field  $C$  such that  $\text{div}(C(x)e^{-U(x)/T}) = 0$  and let  $\Sigma(x)$  be such that  $\Sigma(x) - I$  is nonnegative definite. Then we have

$$\tilde{I}_{f,\Sigma,C}(\ell) \geq \tilde{I}_{f,\Sigma}(\ell) \geq \tilde{I}_{f,o}(\ell),$$

If  $\Sigma(x) - I$  is strictly positive definite, and if there exists  $\ell_0$  such that for this particular field  $C$ ,  $\tilde{I}_{f,\Sigma,C}(\ell_0) = \tilde{I}_{f,\Sigma}(\ell_0)$  or  $\tilde{I}_{f,\Sigma,C}(\ell_0) = \tilde{I}_{f,o}(\ell_0)$  then we must have

$$(19) \quad \hat{\beta}(\ell_0)f = e^{-(G+U)} (\mathcal{L}_0 + \mathcal{S}) e^{G+U},$$

where  $G$  is such that  $G + U$  is invariant under the particular vector field  $C$  and  $\mathcal{L}_0 + \mathcal{S}$  is the infinitesimal generator of the process given in Example 3.

We conclude this section with the proofs of Propositions 2, 3, 4 and Theorem 2.

*Proof of Proposition 2.* Consider the general situation of Theorem 1 with  $a(x) = 2T\Sigma(x)$  and  $b(x) = -\Sigma(x)\nabla U(x)$ . We then have

$$I_{\Sigma}(\mu) = \frac{2T}{8} \int_E \frac{\nabla p(x)\Sigma(x)\nabla p(x)}{p^2(x)} d\mu(x) + \frac{1}{2} \int_E \frac{\nabla p(x)\Sigma(x)\nabla U(x)}{p(x)} d\mu(x) + T \int_E \nabla \phi(x)\Sigma(x)\nabla \phi(x) d\mu(x),$$

where due to reversibility we have  $\phi(x) = \frac{1}{2T}U(x) + \text{constant}$ . Notice that  $I_o(\mu)$  is nothing else but  $I_{\Sigma}(\mu)$  with  $\Sigma(x) = I$ . Taking then, the difference  $I_{\Sigma}(\mu) - I_o(\mu)$  and doing some straightforward algebra, we obtain the statement of the proposition. Clearly,  $I_{\Sigma}(\mu) - I_o(\mu) > 0$  if  $\Sigma(x) - I$  is strictly positive definite.  $\square$

*Proof of Proposition 3.* Considering the general situation of Theorem 1, we obtain for the difference

$$I_{\Sigma,C}(\mu) - I_{\Sigma}(\mu) = \int_E \left[ T(\nabla \phi(x)\Sigma(x)\nabla \phi(x)) - \frac{1}{4T} \nabla U(x)\Sigma(x)\nabla U(x) - \frac{1}{2} \frac{C(x)\nabla p(x)}{p(x)} \right] d\mu(x)$$

Using the condition  $\text{div}(C(x)e^{-U(x)/T}) = 0$ , which can be rewritten as  $\text{div}C(x) = T^{-1}C(x)\nabla U(x)$ , and integrating by parts we get for the last term of the last display

$$\begin{aligned} \int_E \frac{C(x)\nabla p(x)}{p(x)} d\mu(x) &= \int_E C(x)\nabla p(x) dx = - \int_E \text{div}C(x)p(x) dx = - \int_E \text{div}C(x) d\mu(x) \\ &= - \int_E \frac{1}{T} C(x)\nabla U(x) d\mu(x). \end{aligned}$$



Thus we have obtained

$$I_{\Sigma,C}(\mu) - I_{\Sigma}(\mu) = \int_E \left[ T(\nabla\phi(x)\Sigma(x)\nabla\phi(x)) - \frac{1}{4T}\nabla U(x)\Sigma(x)\nabla U(x) + \frac{1}{2T}C(x)\nabla U(x) \right] d\mu(x)$$

Recall now that  $\phi(x)$  is the unique solution, up to constants, of the equation

$$\operatorname{div} [p(x)(-\Sigma(x)\nabla U(x) + C(x) + 2T\Sigma(x)\nabla\phi(x))] = 0.$$

Its weak form reads as follows

$$(20) \quad \int_E \nabla g(x) [2T\Sigma(x)\nabla\phi(x) - \Sigma(x)\nabla U(x) + C(x)] d\mu(x) = 0, \quad \text{for every } g \in \mathcal{C}^1(E)$$

and we can pick freely  $g \in \mathcal{C}^1(E)$ . Let us first choose  $g(x) = \frac{1}{2}\phi(x) + \frac{1}{4T}U(x)$ . Then, (20) gives

$$\int_E \left[ T(\nabla\phi(x)\Sigma(x)\nabla\phi(x)) - \frac{1}{4T}\nabla U(x)\Sigma(x)\nabla U(x) \right] d\mu(x) = - \int_E C(x) \left( \frac{1}{2}\nabla\phi(x) + \frac{1}{4T}\nabla U(x) \right) d\mu(x)$$

and thus, we obtain

$$(21) \quad I_{\Sigma,C}(\mu) - I_{\Sigma}(\mu) = \int_E C(x) \left( -\frac{1}{2}\nabla\phi(x) + \frac{1}{4T}\nabla U(x) \right) d\mu(x)$$

Choosing then  $g(x) = \frac{1}{2}\phi(x) - \frac{1}{4T}U(x)$  and we get from (20) and the latter display

$$I_{\Sigma,C}(\mu) - I_{\Sigma}(\mu) = 4T \int_E \left( \frac{1}{2}\nabla\phi(x) - \frac{1}{4T}\nabla U(x) \right)^T \Sigma(x) \left( \frac{1}{2}\nabla\phi(x) - \frac{1}{4T}\nabla U(x) \right) d\mu(x)$$

which is the statement of the proposition. It is clear that  $I_{\Sigma,C}(\mu) - I_{\Sigma}(\mu) \geq 0$ . If  $\Sigma(x)$  is strictly positive definite and  $\mu$  possesses a strictly positive density, it is clear that  $I_{\Sigma,C}(\mu) - I_{\Sigma}(\mu) = 0$  if and only if  $\operatorname{div}(pC) = 0$ . In other words,  $I_{\Sigma,C}(\mu) - I_{\Sigma}(\mu) > 0$  if and only if  $\operatorname{div}(pC) \neq 0$  and  $\Sigma(x)$  is strictly positive definite. It is clear that if  $\operatorname{div}(p(x)C(x)) = 0$ , then the requirement  $\operatorname{div}(C(x)e^{-U(x)/T}) = 0$  implies that  $p$  can be written as  $p(x) = e^{G(x)}$  with  $C(x)\nabla(G(x) + U(x)) = 0$ .  $\square$

*Proof of Proposition 4.* We write

$$I_{\Sigma,C}(\mu) - I_o(\mu) = [I_{\Sigma}(\mu) - I_o(\mu)] + [I_{\Sigma,C}(\mu) - I_{\Sigma}(\mu)]$$

Notice that the first term on the right hand side of the last display is the difference  $I_{\Sigma}(\mu) - I_o(\mu)$  from Proposition 2, whereas the second term is the difference  $I_{\Sigma,C}(\mu) - I_{\Sigma}(\mu)$  from Proposition 3. This concludes the proof of the proposition.  $\square$

*Proof of Theorem 2.* Analogously to Proposition 4.1 of [31] we have that for each one of the infimization problems defining  $\tilde{I}_{f,\Sigma}(\ell)$ ,  $\tilde{I}_{f,\Sigma,C}(\ell)$  and  $\tilde{I}_{f,o}(\ell)$  there is a corresponding infimizing measure (different for each case)  $\mu^*(dx) = p^*(x)dx$  with  $p^*(x) > 0$  and  $p^*(x) \in \mathcal{C}^{(2+\alpha)}(E)$  that attains the infimum. For example, in the case of only a reversible perturbation we have that

$$\tilde{I}_{f,\Sigma}(\ell) = I_{\Sigma}(\mu^*).$$

Then, a straightforward contradiction argument that is based on Propositions 2, 3 and 4 leads to the proof of the statement  $\tilde{I}_{f,\Sigma,C}(\ell) \geq \tilde{I}_{f,\Sigma}(\ell) \geq \tilde{I}_{f,o}(\ell)$ .

The derivation of the PDE (19) that characterizes the situation where the rate function does not increase goes as follows. It can be seen that  $p^*(x)$  is the invariant density corresponding to the infinitesimal generator  $\mathcal{L} + \nabla\phi_{\hat{\beta}} \cdot \nabla$  where  $e^{\phi_{\hat{\beta}}}$  is the eigenfunction associated to the eigenvalue  $\lambda(\hat{\beta}f)$  defined in (18) for the operator  $\mathcal{L} + \beta f$ . We recall that  $\hat{\beta} = \hat{\beta}(\ell)$  is the unique solution of  $\frac{d}{d\beta}\lambda(\beta f) = \ell$ . Due to the dependence of  $p^*(x)$  on  $\hat{\beta}$ , let us write  $p_{\hat{\beta}}(x)$  for  $p^*(x)$ . In other words,  $p_{\hat{\beta}}(x)$  should satisfy

$$(22) \quad (\mathcal{L} + \nabla\phi_{\hat{\beta}} \cdot \nabla)^* p_{\hat{\beta}} = 0,$$

where  $e^{\phi_{\hat{\beta}}(x)}$  is the eigenfunction associated to the eigenvalue  $\lambda(\hat{\beta}f)$ , i.e.,

$$(23) \quad (\mathcal{L} + \hat{\beta}f)e^{\phi_{\hat{\beta}}} = \lambda(\hat{\beta}f)e^{\phi_{\hat{\beta}}}$$

If the rate function does not increase, then one should have that  $\text{div}(Cp_{\hat{\beta}}) = 0$ . Since  $\text{div}(Cp_{\hat{\beta}}) = 0$  we have that in fact

$$(\mathcal{L}_0 + \mathcal{S} + \nabla\phi_{\hat{\beta}} \cdot \nabla)^* p_{\hat{\beta}} = 0.$$

Since  $\mathcal{L}_0 + \mathcal{S} + \nabla\phi_{\hat{\beta}} \cdot \nabla$  is the generator of a reversible ergodic Markov process, we then obtain that  $p_{\hat{\beta}} = e^{(\phi-U)+\text{const}}$ . Thus,  $\phi = G + U$  and  $C \cdot \nabla\phi = 0$  and (23) reduces to

$$(24) \quad (\mathcal{L}_0 + \mathcal{S} + \hat{\beta}f)e^{\phi_{\hat{\beta}}} = \lambda(\hat{\beta}f)e^{\phi_{\hat{\beta}}}$$

Since changing  $f$  into  $f + c$  leaves  $\phi$  unchanged, but changes  $\lambda(\hat{\beta}f)$  to  $\lambda(\hat{\beta}f) + \hat{\beta}c$ , we get that the equation  $(\mathcal{L}_0 + \mathcal{S} + \hat{\beta}f)e^{\phi_{\hat{\beta}}} = \lambda(\hat{\beta}f)e^{\phi_{\hat{\beta}}}$  implies (19).  $\square$

## 6. CONCLUSIONS

In this paper we have demonstrated in a very general setting that perturbations of the generator of reversible Markov processes by reversible negative definite operators or by irreversible (anti-selfadjoint) operators that maintain the invariant measure lead to improvement in sampling. In particular, we have shown that spectral gap decreases, the asymptotic variance of trajectory time averages decreases and the large deviations rate function that controls the decay rate of the tail distribution of the estimator increases. In all these three cases, we have worked with the generator of the given Markov process. Moreover, we have provided specific reversible and irreversible perturbations for cases of interest such as continuous time Markov chains, Markov jump processes as well as diffusion processes.

Clearly, there are many open questions to address here, perhaps the most important ones being optimal perturbation in different concrete cases of interest, as well as the involved numerical challenges, see [31, 9]. Some preliminary results on optimal perturbations for the case of quadratic  $U(x)$  (the Gaussian case) can be found in [9, 18, 21]. In addition, in most of the cases, what is guaranteed is that ergodic behavior does not become worse. It is interesting to provide concrete conditions for strict improvement, such as the ones for the diffusion case presented in [31] and in Section 5 of the present paper.

## REFERENCES

- [1] S. Asmussen and P.W. Glynn, *Stochastic Simulation*, Springer, 2007.
- [2] K.A. Athreya, H. Doss and J. Sethuraman, On the convergence of the Markov chain simulation method, *Annals of Statistics*, Vol. 24, (1996), pp. 69-100.
- [3] J. Bierkens, Non-reversible Metropolis-Hastings, *Statistics and Computing*, (2015), pp. 1-16.
- [4] T.-L. Chen and C.-R. Hwang, Accelerating reversible Markov chains, *Statistics and Probability Letters*, Vol. 83, Issue 9, (2013), pp. 1956-1962.
- [5] F. Den Hollander, *Large deviations*, American Mathematical Society, Providence, RI, 2000.
- [6] P. Diaconis, S. Holmes and R. Neal, Analysis of a nonreversible Markov chain sampler, *Annals of Applied Probability*, Vol. 10, (2010), pp. 726-752.
- [7] P. Diaconis and L. Miclo, On the spectral analysis of second-order Markov chains, *Annales de la Faculté des Sciences de Toulouse. Mathématiques. Série 6* Vol. 22 (2013), no. 3, 573-621.
- [8] M.D. Donsker and S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large times, I, *Communications Pure in Applied Mathematics*, Vol. 28, (1975), pp. 1-47, II, *Communications on Pure in Applied Mathematics*, Vol. 28, (1975), pp. 279-301, and III, *Communications on Pure in Applied Mathematics*, Vol. 29, (1976), pp. 389-461.
- [9] A. B. Duncan, T. Lelievre, and G. A. Pavliotis, Variance Reduction using Nonreversible Langevin Samplers, *Journal of Statistical Physics*, Vol. 163, Issue 3, (2016), pp 457-491.
- [10] P. Dupuis, Y. Liu, N. Plattner, and J. D. Doll, On the Infinite Swapping Limit for Parallel Tempering. *SIAM Multiscale Modeling and Simulation*, Vol. 10, Issue 3, (2012), pp. 986-1022.
- [11] P. Dupuis and Y. Liu, On the large deviation rate function for the empirical measures of reversible jump Markov processes. *Annals of Probability*, to appear, (2013).
- [12] B. Franke, C.-R. Hwang, H.-M. Pai, and S.-J. Sheu, The behavior of the spectral gap under growing drift, *Transactions of the American Mathematical Society*, Vol 362, No. 3 (2010), pp. 1325-1350.
- [13] A. Frigessi, C.R. Hwang and L. Younes, Optimal spectral structures of reversible stochastic matrices, Monte Carlo methods and the simulation of Markov random fields, *Annals of Applied Probability*, Vol. 2, (1992), pp. 610-628.
- [14] A. Frigessi, C.R. Hwang, S.J. Sheu and P. Di Stefano, Convergence rates of the Gibbs sampler, the Metropolis algorithm, and their single-site updating dynamics, *Journal of Royal Statistical Society Series B, Statistical Methodology*, Vol. 55, (1993), pp. 205-219.
- [15] J. Gärtner, On large deviations from the invariant measure, *Theory of probability and its applications*, Vol. XXII, No. 1, (1977), pp. 24-39.

- [16] M. Girolami and B. Calderhead, Riemann manifold Langevin and Hamiltonian Monte Carlo methods, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 73, Issue 2, (2011), pp. 1232-14.
- [17] E. Hille and P.S. Phillips, *Functional Analysis and Semi-Groups*, American Mathematical Society, Colloquium Publications, Vol.31, 1957.
- [18] C.R. Hwang, S.Y. Hwang-Ma and S.J. Sheu, Accelerating Gaussian diffusions. *The Annals of Applied Probability* Vol. 3, (1993) pp. 897-913.
- [19] C.R. Hwang, S.Y. Hwang-Ma and S.J. Sheu, Accelerating diffusions, *The Annals of Applied Probability*, Vol 15, No. 2, (2005), pp. 1433-1444.
- [20] C.-R. Hwang, R. Normand and S.-J. Wu, Variance reduction for diffusions, *Stochastic Processes and their Applications*, Vol. 125, No. 9, (2015), pp. 3522-3540.
- [21] T. Lelièvre, F. Nier and G.A. Pavliotis, Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion, *Journal of Statistical Physics*, Vol. 152, Issue 2, 237-274, (2013)
- [22] F. Leisen and A. Mira, An extension of Peskun and Tierney orderings to continuous time Markov chains, *Statistica Sinica*, Vol. 18, (2008), pp. 1641-1651.
- [23] C. Maes, K. Netočný and B. Wynants, Monotonicity of the dynamical activity, *J. Phys. A: Math. Theor.* Vol. 45, (2012) pp. 455001.
- [24] K.L. Mengersen and R.L. Tweedie, Rates of convergence of the Hastings and Metropolis algorithms, *Annals of Statistics*, Vol. 24, (1996), pp. 101-121.
- [25] S. Meyn and R.L. Tweedie, *Markov Chains and Stochastic Stability*, Cambridge University Press, Second Edition, 2009.
- [26] A. Mira, Efficiency of finite state space Monte Carlo Markov chains, *Statist. Probab. Lett.*, Vol. 54, No. 4, (2001), pp. 405-411.
- [27] A. Mira, Ordering and improving the performance of Monte Carlo Markov chains, *Statist. Sci.*, Vol. 16, No. 4, (2001), pp. 340-350.
- [28] A. Mira and C. J. Geyer, On non-reversible Markov chains, In *Monte Carlo methods, Volume 26 of Fields Inst. Commun.*, Amer. Math. Soc., Providence, RI, (2000), pp. 95-110.
- [29] R.M. Neal, Improving asymptotic variance of MCMC estimators: Non-reversible chains are better, *Technical report, No. 0406, Department of Statistics, University of Toronto*, 2004.
- [30] P. H. Peskun, Optimum Monte-Carlo sampling using Markov chains, *Biometrika*, Vol. 60, (1973), pp. 607-612.
- [31] L. Rey-Bellet and K. Spiliopoulos, Irreversible Langevin samplers and variance reduction: a large deviations approach, *Nonlinearity*, Vol. 28, (2015), pp. 2081-2103.
- [32] L. Rey-Bellet and K. Spiliopoulos, Variance reduction for irreversible Langevin samplers and diffusion on graphs, *Electronic Communications in Probability*, Vol. 20, no. 15, (2015), pp. 1-16.
- [33] G.O. Roberts and J.S. Rosenthal, General state space Markov Chain and MCMC algorithms, *Probability Surveys*, Vol. 1, (2004), pp. 20-71.
- [34] F. Schlögl, Chemical reaction models for nonequilibrium phase transition, *Z. Physik*, Vol. 253, (1972), pp. 147-161.
- [35] Y. Sun, F. Gomez, and J. Schmidhuber, Improving the Asymptotic Performance of Markov Chain Monte- Carlo by Inserting Vortices. In *Advances in Neural Information Processing Systems* Vol. 23, (2010), pp. 2235-2243.
- [36] H. Suwa, and S. Todo, Markov Chain Monte Carlo Method without Detailed Balance, *Phys. Rev. Lett.*, Vol. 105, (2010), pp. 120603.
- [37] L. Tierney, A note on Metropolis-Hastings kernels for general state spaces, *The Annals of Applied Probability*, Vol. 8, No. 1, (1998), pp. 1-9.
- [38] K.S. Turitsyn, M. Chetkov, and M. Vucelja, Irreversible Monte Carlo Algorithms for Efficient Sampling, *Physica D*, Vol. 240, (2011), pp. 410.
- [39] A. Ichiki, and M. Ohzeki, Violation of detailed balance accelerates relaxation, *Phys. Rev. E*, Vol. 88, (2013), pp. 020101(R).